Reconhecimento Off-line de Voz Contínuo para Dispositivos Móveis: Uma Análise Comparativa de Métricas de Avaliação

Lucas Debatin[1, 2], Aluizio Haendchen Filho[1, 2], Rudimar L. S. Dazzi[1]

[1] Laboratório de Inteligência Aplicada – Universidade do Vale do Itajaí (UNIVALI)
Caixa Postal 360 – 88302-202 – Itajaí – SC – Brasil

[2] Núcleo de Inteligência Artificial e Sistemas Inteligentes – Centro Universitário de Brusque (UNIFEBE) – Caixa Postal 1501 – 88352-400 – Brusque – SC – Brasil

lucasdebatin@edu.univali.br, {aluizio.h.filho,rudimar}@univali.br

Abstract. *Speech recognition is a form of accessibility used to perform tasks with hands and eyes free, and this is advantageous regardless of the user type. Current APIs make implementation easy, but they have limitations because depend on Internet connectivity and, are often proprietary software. The proposed solution to resolve these limitations is the development of an off-line continuous speech recognition system. The best techniques selected in a systematic review were implemented using libraries. This work presents a comparative analysis of the evaluation metrics obtained for each library.*

1. Introdução

O reconhecimento de voz é o processo que converte a linguagem falada em texto, e pode ser classificado em: (*i*) palavras isoladas, que necessita de uma pausa maior que 200ms entre as palavras da frase; e (*ii*) contínuo, que reconhece sentenças pronunciadas de forma natural, sem pausas entre as palavras [Alencar 2005; Huang e Deng 2010; Silva 2010].

Atualmente, as APIs (*Application Programming Interface*) para implementação do reconhecimento de voz apresentam limitações, tais como: (*i*) nenhuma realiza o reconhecimento em modo off-line; e (*ii*) são softwares proprietários, e em muitos casos o valor pago pela licença de uso pode se tornar alto, dependendo da quantidade de requisições [Debatin et al. 2018]. O termo off-line se refere a ausência de conexão com a internet ao utilizar o reconhecimento de voz contínuo em softwares e aplicativos.

Nesse trabalho serão apresentados os resultados da implementação das bibliotecas CMUSphinx, HTK e Kaldi de reconhecimento de voz em um computador desktop.

2. Implementação das Bibliotecas

Com base nos resultados obtidos em [Debatin et al. 2018], os seguintes passos foram executados para a implementação: (*i*) extração de características do áudio com a técnica MFCC *(Mel-Frequency Cepstral Coefficient*); (*ii*) modelo acústico composto por um dicionário fonético e as redes DNNs (*Deep Neural Networks*) e/ou o HMM (*Hidden Markov Model*); e (*iii*) modelo de linguagem que utiliza técnicas de bigrama e trigrama. Os modelos são responsáveis por fornecer a melhor representação textual com base nas características extraídas do áudio de entrada.

Para a implementação, utilizou-se as bibliotecas [Matarneh et al. 2017]: (*i*) CMUSphinx que adota o modelo de linguagem estatística HMM e n-grama, capaz de

realizar o reconhecimento contínuo com grande vocabulário; (*ii*) HTK que utiliza algoritmos clássicos de reconhecimento de voz e estruturas de dados, com o modelo de linguagem estatística HMM e n-grama; e (*iii*) Kaldi que permite usar uma variedade de algoritmos para aumentar o desempenho do sistema por meio da redução do tamanho das características do sinal acústico.

3. Resultados Iniciais

A Tabela 1, apresenta as melhores métricas de avaliação, WER (*Word Error Rate*) e SER (*Sentence Error Rate*), obtidas pelas três bibliotecas e para cada corpus de voz. Para o treinamento e testes, utilizou-se 90% e 10%, respectivamente, do tamanho dos corpora de voz do grupo FalaBrasil. Dentre as várias abordagens que a biblioteca Kaldi fornece, a MLP-HMM (*Multilayer Perceptron-Hidden Markov Model*) obteve o melhor resultado.

Tabela 1. Métricas de avaliação obtidas pelas bibliotecas para cada corpus de voz

Corpus/Biblioteca	CMUSphinx	HTK	Kaldi
LaPSBenchmark (54 minutos de áudio de vários locutores)	WER: 9,1% SER: 56,7%	WER: 94,95% SER: 100%	WER: 4,23% SER: 23,33%
Constituição Federal (9 horas de áudio de um locutor)	WER: 3,1% SER: 73%	WER: 81,72% SER: 100%	WER: 0,83% SER: 34,92%

Os resultados mostram que o uso da biblioteca Kaldi obteve as melhores métricas de avaliação (WER e SER). O principal motivo deve-se ao fato de que essa biblioteca é a única dentre as três que apresenta o uso de DNNs.

4. Conclusão e Trabalhos Futuros

Esse trabalho contribuiu para a pesquisa, na área de reconhecimento de voz contínuo, por meio de uma análise comparativa das melhores métricas de avaliação obtidas com a implementação das bibliotecas.

Os trabalhos futuros estão direcionados para a implementação das bibliotecas em dispositivos móveis com os sistemas operacionais IOS e Android, com o intuito de realizar testes para averiguar a performance e otimização do uso de memória.

Referências

Alencar, V. F. S. (2005) “Atributos e Domínios de Interpolação Eficientes em Reconhecimento de Voz Distribuído”, In: PUC-Rio, Rio de Janeiro, Brazil.

Debatin, L., Haendchen Filho, A. and Dazzi, R. L. S. (2018). Offline Speech Recognition Development: A Systematic Review of the Literature. In *International Conference on Enterprise Information Systems*, p. 551-558.

Huang, X. and Deng, L. (2010) “An overview of modern speech recognition”, In: Handbook of Natural Language Processing, Edited by Nitin Indurkhya, Fred J. Damerau, 2th edition. London: Chapman and Hall/CRC, p. 339-366.

Matarneh, R., Maksymova, S., Lyashenko, V. V. and Belova, N. V. (2017). Speech Recognition Systems: A Comparative Review. In *IOSR Journal of Computer Engineering*, p. 71-79.

Silva, C. P. A. (2010) “Um software de reconhecimento de voz para português brasileiro”, In: Universidade Federal do Pará, Pará, Brazil.